METHODS AND SYSTEMS FOR ENABLING EFFICIENT SEARCH AND RETRIEVAL OF RECORDS FROM A COLLECTION OF BIOLOGICAL DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority to and incorporates by reference in its entirety provisional application serial no. 09/193,263, filed March 30, 2000 entitled "METHODS AND SYSTEMS FOR ENABLING REVENUE MODELS BASED ON THE INSTANTANEOUS PREFERENCES OF ON-LINE USERS".

BACKGROUND OF THE INVENTION

Field Of The Invention

The present invention relates to systems and methods for searching a collection of biological data in such a manner that it is easy to search, drill down, drill-up and drill across records in the bioinformatics data collection using multiple, independent hierarchical category taxonomies of the records in the bioinformatics data collection.

Description of the Related Art

The present invention is directed to systems and methods for quickly and efficiently retrieving information from a bioinformatics data collection.

Recent advances in life science research have dramatically increased the rate at which information is being produced. This data, which is continuously analyzed by researchers, is stored in databases hosted at various institutions throughout the world. There are hundreds of these

1

databases that hold information regarding the human genome, proteomics, biological pathways and processes. The first step a scientist often takes during the course of conducting research is to consult these databases to see what findings exist that may be similar or helpful in their research. This information is stored in a traditional database with an input box front end for the researchers

5   to type in their criteria and keywords. The amount of data is growing exponentially. The results that come back are very poor. For these researchers and their corporations, speed is everything. Speed to research, speed to patent, speed to drug discovery, etc.

With the dramatic increase in the amount of data that exists, and the increased speed with which it needs to be analyzed, however, has come the need for better ways in which to navigate

10   electronically stored information. Historically, a few of the fields are filled out in the database input box and the annotations come back in a long list. There is no option to browse or discover. In parallel, ontology schemes are being developed to overlay this wealth of life sciences information, to be better able to communicate and analyze it.

There is a need, therefore, for overcoming the inherent deficiencies in utilizing search

15   engines to navigate vast numbers of electronically stored biological records. There is a need to ensure that a search engine yields a list of records that are significantly relevant to the search expression provided by the user. That is, there is a need for an engine that yields greater accuracy in performing a search of electronically stored biological records for only those records related to a given search expression.

20   Figure 1 is a visual representation of a bioinformatics data collection 1. This bioinformatics data collection 1 is made up of a plurality of records of biological data 2. Each record of biological data may consist of a single character, a string of characters, a plurality of strings of characters, an image, an audio file or any combination of the preceding. The size of the

2

bioinformatics data collection 1 can be described by making reference to the number of records of biological data 2 within it. Large bioinformatic data collections may contain millions or billions of records regarding biological data.

The task of a bioinformatics data collection search engine is to provide the user with a list of biological data that the search engine calculates is likely to hold information chosen by the user. This list is compounded by using a search term or query 3. One method of compounding this list is a full-text algorithm. A "full-text" search algorithm identifies biological records that contain key term(s) in each and every record of biological data. In other words, the search process effectively identifies records such as record 2 that contain the search term 3. When the search is completed, a numerical count of the total number of records for biological data containing the search term(s) is compiled and displayed along with a list of links to those biological data to allow the user to view the biological data. That is, the number of matches, e.g., "2,000 matches," links and descriptions of the first few matching biological records are displayed to the user. The user reviews the number of matches and the provided descriptions of some of the matched biological records and either decides to try a different search in an attempt to shrink the number of matches or selects one listed link to access a particular record.

One problem with these types of search engines is the often-large number of matches returned to the user. If a user enters the search term "cell," he/she may receive over 1 million matches. Almost no user will wade through all 1 million biological records looking for the best or specific record that he/she needs.

If the user edits the search term(s), he/she may pare the number of matches down from 1 million to 200,000, but this number of matches is still too large for a user to view and use to make an effective decision. The user may then try to re-edit the search terms in an iterative process

3

until the number of matches is manageable. However, this iterative process of re-editing search terms is time consuming and may frustrate the user before he/she receives the desired data.

In an effort to reduce this frustration, search engines were developed that categorize the records and provide the categories to the user so that he/she may reduce the number of records

5   before executing a search using search term(s).

Figure 2 shows some records 205, 210 and 215 from bioinformatics data collection 1. These records are categorized. The exemplary categories 250 shown are "Cell Communication," "Cell Adhesion" and "Flocculationi;" "Cell Growth & Maintenance," "Cell Cycle" and "Nuclear Migration;" and "Developmental Processes," "Gametogenesis" and "Oogenesis." These

10   categories 250 relate to the taxonomy "Biological Processes."

One method of categorizing records of biological data is to apply tags to each record. For example, if biological data contains records which relate to a certain type, then that record is tagged with a unique tag identifying its relationship to that type. Other records that do not contain data related to that type are not tagged with that unique tag. These tags are later used to identify

15   and retrieve records of biological data containing data related to certain types. As a further example, if a record contains the word "plasma," then that record is tagged with a tag called "PL."

The categorized records of biological data 205, 210 and 215 are tagged with a single taxonomy because all of the categories 250 represent a class or subset of the taxonomy "Biological Processes." Assuming all of the records of biological data within bioinformatics data

20   collection 1 are categorized, bioinformatics data collection 1 can be referred to as a "single-taxonomy, categorized bioinformatics data collection."

Given these definitions, it is clear that a taxonomy is a hierarchical organization of categories and the various taxonomies and categories inherent to a bioinformatic record can be

used to organize the records of biological data in a bioinformatics data collection. This organization of the records of the biological data, in turn, makes it easier to search for, retrieve, and display records containing specific data. In other words, a user may use the taxonomies and categories to search bioinformatics data collection 1 if the records in bioinformatics data

5   collection 1 are properly tagged.

Typically, taxonomies and categories are selected from among those characteristics and attributes which a user would intuitively think of to launch a search. For instance, a user attempting to find fibrillar collagen genes would formulate a search based on certain intuitive characteristics, one being the "molecular function" of genetic records in bioinformatics data

10   collection 1. This intuitive characteristic becomes a taxonomy. This search can be narrowed by using the attributes "Extracellular", "Extracellular Matrix" and "Collagen." These intuitive attributes are categories within the taxonomy.

One problem with most conventional search tools based on categories is that they only provide the user with a single taxonomy. For example, assume that a user searches using a

15   taxonomy called "Molecular Function" and a category called "Signal Transduction" to identify all related "Ligand" genes. Suppose now, however, the user wishes to identify only those "Ligand" genes with a biological process of "Behavior". For a single taxonomy-categorized search, this means launching a new search because "Behavior" is neither an attribute nor a characteristic related to "Molecular Function." Instead, "Behavior" is independent of record type and is related

20   to a different taxonomy, such as "Biological Process."

To try to alleviate this problem, many single-taxonomy, categorized search engines allow Boolean operations. Thus, if the user discovers that there are 100 different records of biological data, he/she may further refine this search by searching for the word "Behavior." Thus, the user

edits the search to be "Ligand" AND "Behavior." This type of search modification is only marginally effective, for several reasons. First, the use of a Boolean search at this point usually entails the initiation of a new search. Second, the search engine, because it does not provide a taxonomy, cannot suggest terms for narrowing the search to the desired data, which requires the user to be clear about and know the Boolean query terms in advance.

In an attempt to address data searching of ever increasing bioinformatics data collections, many techniques have been developed. For example, U.S. Patent Number 5,675,786 relates to accessing data held in large computer databases by sampling the initial result of a query of the database. Sampling of the initial result is achieved by setting a sampling rate which corresponds to the intended ratio at which the data documents of the initial result are to be sampled. The sampling result is substantially smaller than the initial query result and is thus easier to analyze statistically. While this method decreases the amount of data sent as a result of the query to the end user, it still results in an initial search of what could be a massive database. Further, dependent upon the sampling rate, sampling may result in a reduction in the accuracy of the information sent to the end user and may thus not provide the intended result.

Another example, U.S. Patent Number 5,642,502 relates to a method and system for searching and retrieving documents in a database. A first search and retrieval result is compiled on the basis of a query. Each word in both the query and the search result are given a weighted value, and then combined to produce a similarity value for each document. Each document is ranked according to the similarity value and the end user chooses documents from the ranking. On the basis of the documents chosen from the ranking, the original query is updated in a second search and a second group of documents is produced. The second group of documents is supposed to have the more relevant documents of the query closer to the top of the list. While more relevant

documents may be found as a result of the second search, the patent does not address the problems

associated with the searching of a large database and, in fact, might only compound them.

Additionally, the patent does not return categorized search results complete with counts of the

number of records associated with those categories.

5   Yet another example, U.S. Patent Number 5,265,244 relates to a method and apparatus for

data access using a particular data structure. The structure has a plurality of data nodes, each for

storing data, and a plurality of access nodes, each for pointing to another access node or a data

node. Information, of a statistical nature, is associated with a subset of the access nodes and data

nodes in which the statistical information is stored. Thus statistical information can be retrieved

10   using statistical queries which isolate the subset of the access nodes and data nodes which contain

the statistical information. While the patent may save time in terms of access to the statistical

information, user access to the actual data documents requires further procedures.

Further, U.S. Patent No. 5,930,474 discloses a search engine configured to search

geographically and topically, wherein the search engine is configurable to search for user-entered

15   topics within a hierarchically specified geographic area. This system makes use of a static index

of results for each taxonomy. Because this system does not produce dynamic search results, it

precludes the ability to switch among multiple taxonomies. The system is also not text searchable

at any time during a drill-down. The system also doesn't include counts of records with category

results.

20   U.S. Patent No. 6,012,055 discloses a search system comprising multiple navigators

switchable by tabs in the GUI, having the ability to cross-reference amongst said navigators. This

is just a method for accessing different information sources, not a method for text-searching.

Further, it does not offer user-categorized search results with counts.

7

U.S. Patent No. 5,682,525 discloses an online directory, having the capability to display an advertisement incorporated within a map display, wherein the said map has indicia for points of interests selected by a user from a drop down menu. This invention describes a technique for identifying targeted advertising based on categories selected within a hierarchical taxonomy. This

5   invention does not consider cross-sections of categories across multiple taxonomies, i.e. biological process, molecular function and cellular component. Nor does this invention consider the addition of keyword searches as a further limiting item for identifying targeted advertising. U.S. Patent No. 6,078,916 discloses a search engine which displays an advertising banner having a keyword associated therewith, wherein the keyword is related to a user-entered search topic.

10   This invention discloses a method for organizing information based on the statistics and heuristical information derived from a user's behavior.

Megaspider, a meta-search engine, has a web directory with hierarchically arranged geographic regions, having subcategories therein for topics, said directory being searchable within a geographic area or within a topic. However, MegaSpider's search technology employs a static

15   hierarchical drill-down and cannot execute a full-text search and return categorized search results with counts. Additionally, this system only has one hierarchical taxonomy and cannot switch between multiple taxonomies, nor yield categorized search results with counts when searching.

U.S. Patent No. 5,832,497 discloses a system which enables users to search for jobs by geographical location and specialty. While this invention does discuss an iterative method for

20   finding information in a multi-dimensional database, it does not consider categorized search results with counts (i.e. the ability to conduct a field or free-text search and have the results be returned by one or many sets of hierarchically organized categories with counts of the number of records associated with each of those categories), nor the ability to switch among taxonomies.

However, none of these conventional systems provide users with a multiple-taxonomy, multiple-category search engine that allows users to search for documents, where the user is allowed to toggle among the multiple taxonomies as an aid to locating desired documents without constraints.

5

SUMMARY OF THE INVENTION

The present invention overcomes the shortcomings identified above. More specifically, the present invention is a multiple-taxonomy, multiple category search tool that allows a user to "navigate" through a bioinformatics data collection using any of the taxonomies at any time.

10 In addition, the present invention overcomes the identified shortcomings of other search engines when small screen devices are employed to display search results. More specifically, the present invention transmits and displays categories for users to select from rather than providing users with long laundry lists of record hits.

Through the presentation of categorized search results, the present invention allows an 15 enormous database to be represented by a very small footprint, which is ideal for wireless devices.

Further, the present invention provides a mechanism for "slicing-and-dicing" the information in a database, thus allowing the creation of personalized or customized data collections of bioinformatic data.

The present invention provides such advantages by means of a system for searching a 20 collection of data, said system comprising: an organizer configured to receive search requests, said organizer comprising: a collection of data having at least two entries; wherein the collection of data is organized into at least two taxonomies; wherein each of the at least two taxonomies is associated with at least two categories; wherein the entries correspond to at least one of the at

least two taxonomies and also correspond to at least one of the at least two categories; and a search engine in communication with the collection of data, wherein said search engine is configured to search based on the at least two taxonomies and based on the at least two categories, wherein the search engine returns, in response to a search request identifying at least a first

5   taxonomy of the at least two taxonomies, a list of the categories associated with the at least first identified taxonomy, along with the number of entries associated with each of the categories associated with the at least first identified taxonomy.

The above advantages are further provided through the present invention, which is a system for searching a collection of data, said system comprising: means for networking a

10   plurality of computers; and means for organizing executing in said computer network and configured to receive search requests from any one of said plurality of computers, said means for organizing comprising: a collection of data having at least two entries; wherein the collection of data is organized into at least two taxonomies; wherein each of the at least two taxonomies is associated with at least two categories; wherein the entries correspond to at least one of the at

15   least two taxonomies and also correspond to at least one of the at least two categories; and means for searching in communication with the collection of data, wherein said means for searching is configured to search based on the at least two taxonomies and based on the at least two categories, wherein the means for searching returns, in response to a search request identifying one of the at least two taxonomies, a list of the categories associated with the identified taxonomies, along with

20   the number of entries associated with each of the categories associated with the identified taxonomies.

The above-identified advantages are further provided through a system for searching a collection of data, said system comprising: means for networking a plurality of computers; and

10

means for organizing executing in said computer network and configured to receive search requests from any one of said plurality of computers, said means for organizing comprising: a collection of data having at least two entries; wherein the collection of data is organized into at least two taxonomies; wherein each of the at least two taxonomies is associated with at least two

5 categories; wherein the entries correspond to at least one of the at least two taxonomies and also correspond to at least one of the at least two categories; and means for searching in communication with the collection of data, wherein said means for searching is configured to search based on the at least two taxonomies and based on the at least two categories, wherein the means for searching returns, in response to a search request identifying one of the at least two

10 taxonomies, a list of the categories associated with the identified taxonomy, along with the number of entries associated with each of the categories associated with the identified taxonomy.

Additionally, the above-identified advantages are provided through an article of manufacture comprising: a computer usable medium having computer program code means embodied thereon for searching a collection of data, the computer readable program code means

15 in said article of manufacture comprising: computer readable program code means for communicating a search request to a search engine, the search engine being in communication with a collection of data; wherein the collection of data has at least two entries; wherein the collection of data is organized into at least two taxonomies; wherein each of the at least two taxonomies is associated with at least two categories; wherein the at least two entries correspond

20 to at least one of the at least two taxonomies and also correspond to at least one of the at least two categories; computer readable program code means for querying of the collection of data by the search engine based on the communicated search request; wherein a communicated search request identifies at least one of the at least two taxonomies; and computer readable program code means

11

for returning of a list of the categories associated with the at least one identified taxonomy, along with the number of entries associated with each of the categories associated with the at least one identified taxonomy as a response to the querying of the collection of data.

When potential users navigate a bioinformatics data collection powered by the present search technology, they are greeted with an "aerial" view of the entire bioinformatics data collection. Users thus have the ability to intuitively navigate through huge amounts of information by using keywords and categories in conjunction with the different taxonomies of the bioinformatics data collection. These navigation features are a significant aspect of this bioinformatics data collection search that differentiates it from conventional search technology.

When a user knows what he/she is looking for, the invention quickly uncovers the right information without forcing the user to go through numerous irrelevant search results. The real power of the search technology comes when users do not know or are only vaguely familiar with what they want. In these instances, where a user needs to browse through all or part of the bioinformatic records, keyword searches with categorized search results (from different taxonomies) will facilitate easy navigation by providing the user with context and scope relating to the search results and by giving a user the information he/she needs to find the records of biological data and information he/she required.

The present invention provides users with an aerial view of the bioinformatics data collection at all times during a search. Users remain aware of where they stand in their search and how many records potentially satisfy their query. More importantly, users receive categorized search results that provide summary information on the records in the bioinformatics data collection that remain within the parameters of a search.

Users of the present invention can look for information using keywords they feel will help them refine their search. The system will locate every record in the bioinformatics data collection that contains that particular character-string and instantly return all the record categories (at the category level of the search as then being conducted) that have associated biological data. The search results indicate how many records exist within each applicable category, and allow users to easily hone down on the specific segment of the bioinformatics data collection he/she is interested in and, more importantly, to disregard all other irrelevant information.

For example, if a user enters the search term "acid," the system would search all the records in the bioinformatics data collection that contained the character-string "acid." Rather than returning a long list of numerous search results that satisfy the user's query, the present invention provides the user with the categories that are associated with the remaining records and indicates how many records exist under each category. This functionality assists the user to further refine his/her search and disregard the irrelevant information.

These searched data collections provide users with summary information (categorized search results) about the data collection being searched. Users need not use pull-down menus or fill in any "required" fields to construct the parameters of their search (biological process, molecular function, cellular component, organism, etc.). Rather, search results display only the valid categories and indicate how many records are associated with each applicable category. Users are thus presented with the available options in the bioinformatics data collection (through a dynamic aisle and shelf structure) and can drill down through hierarchically organized bioinformatics data collection or switch among taxonomies to find what they require.

In instances where data collection information can be associated with more than one independent category structure (e.g., biological process, molecular function and cellular

13

component), users of the present invention can switch among taxonomies of the bioinformatic

data collection at any time during the search process and look at information from different

perspectives. Users thus have the ability to navigate through a bioinformatics data collection

using categorized search results that are provided from several different perspectives, or

5   taxonomies. Amazingly, the whole process is extremely intuitive and very easy to use. By using

keywords in conjunction with the different taxonomies of a bioinformatics data collection and by

drilling down hierarchical categories within each taxonomy, users are always left with a refined

set of listings – without having to go through irrelevant search results.

      If a user is drilling down the "Biological Process" taxonomy and clicks on the "Molecular

10   Function" tab, the present invention will instantly reorganize all the records that remain within the

parameters of the search (regardless of number) and present the same information categorized by a

"Molecular Function" taxonomy of the bioinformatics data collection. Switching among

taxonomies is possible at any point in the search process.

15   BRIEF DESCRIPTION OF THE DRAWINGS

      Figure 1 is a simplified diagram of a bioinformatics data collection;

      Figure 2 is a simplified view of various records;

      Figure 3 is a system in accordance with a preferred embodiment of the present invention;

      Figures 4-6 are screen shots a user would see when using an embodiment of the present

20   invention as applied to a biological database;

      Figure 7 is a representation of how a query interacts with indices and how those indices

relate to records of biological data in a bioinformatics data collection according to an embodiment

of the present invention;

Figures 8-10 represent process steps a user would go through to drill down to a set of records in a collection of biological data, in accordance with an embodiment of the present invention;

Figure 11 is a system in accordance with a preferred embodiment of the present invention;

Figure 12 shows a searching process in accordance with an embodiment of the present invention;

Figure 13 is a screen shot of a categorizer in accordance with an embodiment of the present invention;

Figure 14 is a representation of categories and reads in accordance with an embodiment of the present invention;

Figure 15 illustrates a method of distributing, indexing and retrieving data in a distributed data retrieval system, according to an embodiment of the present invention;

Figure 16 illustrates the distribution of data information and the formation of sub-collections in a distributed data retrieval system, according to an embodiment of the present invention;

Figure 17 illustrates an inverted index from which a sub-collection view can be generated in a distributed data retrieval system, according to an embodiment of the present invention;

Figure 18 illustrates a sub-collection view, according to an embodiment of the present invention;

Figure 19 illustrates the paths of communication forming a network between a central computer and a series of local computers in a distributed data retrieval system, according to an embodiment of the present invention; and

Figure 20 illustrates a global view, according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

On-line computer services, such as the Internet, have grown immensely in popularity over

the last decade. Such an on-line computer service can provide access to a hierarchically structured

5   bioinformatics data collection where information within the bioinformatics data collection is

accessible at a plurality of computer servers which are in communication via conventional

telephone lines or T1 links, and a network backbone. For example, the Internet is a giant

internetwork created originally by linking various research and defense networks (such as NSFnet,

MILnet, and CREN). Since the origin of the Internet, various other private and public networks

10   have become attached to the Internet.

The structure of the Internet is a network backbone with networks branching off of the

backbone. These branches, in turn, have networks branching off of them, and so on. Routers move

information packets between network levels, and then from network to network, until the packet

reaches the neighborhood of its destination. From the destination, the destination network's host

15   directs the information packet to the appropriate terminal, or node. For a more detailed description

of the structure and operation of the Internet, please refer to "The Internet Complete Reference,"

by Harley Hahn and Rick Stout, published by McGraw-Hill, 1994.

A user may access the Internet, for example, using a home personal computer (PC)

equipped with a conventional modem. Special interface software is installed within the PC so that

20   when the user wishes to access the Internet, a modem within the user's PC is automatically

instructed to dial the telephone number associated with the local Internet host server. The user can

then access information at any address accessible over the Internet. One well-known software

16

interface, for example, is the Microsoft Internet Explorer (a species of HTTP Browser), developed by Microsoft.

Information exchanged over the Internet is often encoded in HyperText Mark-up Language (HTML) format. HTML encoding is a kind of markup language which is used to define record

5   content information. As is well known in the art, HTML is a set of conventions for marking portions of a record so that, when accessed by a parser, each portion appears with a distinctive format. The HTML indicates, or "tags," what portion of the record the text corresponds to (e.g., the title, header, body text, etc.), and the parser actually formats the record in the specified manner. An HTML document sometimes includes hyper-links which allow a user to move from

10   document to document on the Internet. A hyper-link is an underlined or otherwise emphasized portion of text or graphical image which, when clicked using a mouse, activates a software connection module which allows the users to jump between documents (i.e., within the same Internet site (address) or at other Internet sites). Hyper-links are well known in the art.

One popular computer on-line service is the Web which constitutes a subnetwork of on-

15   line documents within the Internet. The Web includes graphics files in addition to text files and other information which can be accessed using a network browser which serves as a graphical interface between the on-line Web documents and the user. One such popular browser is the MOSAIC web browser (developed by the National Super Computer Agency (NSCA)). A web browser is a software interface which serves as a text and/or graphics link between the user's

20   terminal and the Internet networked documents. Thus, a web browser allows the user to "visit" multiple web sites on the Internet.

Typically, a web site is defined by an Internet address which has an associated home page. Generally, multiple subdirectories can be accessed from a home page. While in a given home

17

page, a user is typically given access only to subdirectories within the home page site; however, hyper-links allow a user to access other home pages, or subdirectories of other home pages, while remaining linked to the current home page in which the user is browsing.

Although the Internet, together with other on-line computer services, has been used widely as a means of sharing information amongst a plurality of users, current Internet browsers and other interfaces have suffered from a number of shortcomings. For example, the organization of information accessible through current Internet browsers and organizers such as Microsoft Internet Explorer or MOSAIC, may not be suitable for a number of desirable applications. In certain instances, a user may desire to access information predicated upon record type as opposed to by subject matter or keyword searches. In addition, present Internet organizers do not effectively integrate record-related information in a consistent manner.

In addition, given the large volume of information available over the Internet, current systems may not be flexible enough to provide for organization and display of each of the kinds of information available over the Internet in a manner which is appropriate for the amount and kind of data to be displayed.

Figure 3 is a system overview in accordance with a preferred embodiment of the present invention. A plurality of user computers 3, 3a and 3b are coupled to a network 2. Network 2 is also coupled to another network 2a which itself is coupled to other computers (not shown). Computer 10 is also coupled to network 2. Coupled to computer 10 is bioinformatics data collection 1. Bioinformatics data collection 1 contains a plurality of records (not shown).

The network 2 may be a private or public network, an intranet or Internet, or a wide or local area network which not only connects the user 3 but other users 3a, 3b and other networks 2a to computer 10.

18

For ease of understanding, in the discussion which follows, the network 2 will comprise the Internet, though this need not be the case.

It should be understood that bioinformatics data collection 1 comprises a multiple-taxonomy, categorized bioinformatics data collection. In such a bioinformatics data collection the records have been tagged or otherwise categorized by more than one taxonomy. For example, the records in bioinformatics data collection 1 have been categorized by the taxonomies "Biological Process," "Molecular Function" and "Cellular Component."

Each taxonomy, in turn, comprises a number of categories. To distinguish the categories and taxonomies used to tag records within bioinformatics data collection 1 from those selected by the user, the categories and taxonomies used to tag the records will be referred to as "database categories" and "database taxonomies."

In one embodiment of the invention, computer 10 receives search requests in the form of data (hereafter referred to as "search-related data") via network 2 from user computer 3. Search-related data comprise a search term entered by a user to initiate a keyword search, or a taxonomy or category selected by the user by "clicking on" a portion of a screen.

The category and/or taxonomy selected by the user and sent to computer 10 is a way for the user to navigate a Web site. As such, the category will be referred to as a "navigational category" and the taxonomy will be referred to as a "navigational taxonomy."

For example, when the user accesses a web site, like web site 4000a and 4000b in Figure 4, he/she is presented with an initial screen which displays taxonomies 4001, 4002 and 4003, namely "Biological Process" 4001, "Molecular Function" 4002 and "Cellular Component" 4003. The user is also presented with organism search scope parameters 4004, 4005 and 4006 with which all the genetic records are associated, namely "Mus" (Mouse) 4004, "Saccharomyces"

19

(Yeast) 4005 and "Drosophila" (Fruit Fly) 4006. In this example the user has decided not to limit the search pool and has selected all three scope parameters 4004, 4005 and 4006. However, in an alternative example the user could have unchecked one or two of these scope parameters and removed from the search pool the number of genetic records associated with each scope

5   parameter.

In this example, the user selects the "Biological Process" taxonomy 4001. After selecting a taxonomy, the user then selects a category 502.

Once computer 10 receives the search-related data, the present invention utilizes the navigational taxonomy 4002 and category 502 in the user's search request to determine sub-

10   categories from the hierarchy associated with the navigational taxonomy and category.

For instance, if the category 502 comprises "Cell Growth and Maintenance," then the process might yield sub-categories 503 shown in Figure 4000b. One such sub-category 503 is "Oncogenesis" 504. Sub-categories 503 will be referred to as "navigational sub-categories."

Once computer 10 has determined the sub-categories 503, it then can launch a search

15   directed to bioinformatics data collection 1.

It will be appreciated that the present invention envisions computer 10 launching search queries aimed at bioinformatics data collection 1 using sub-categories 503 which are not selected by the user. Rather, these sub-categories are dynamically selected by computer 10 based on the taxonomies and/or categories input by the user.

20   According to one embodiment of the present invention, a search query may be carried out in a number of ways.

For example, in one illustrative embodiment of the present invention computer 10 launches a search query comprising a search term 3001, a taxonomy 4001 and sub-categories 503

directed to bioinformatics data collection 1. Computer 10 compares the navigational taxonomy

and sub-categories 503 to the record taxonomies and sub-categories making up bioinformatics

data collection 1. If a record is tagged with a biological data taxonomy and a sub-category which

matches a navigational taxonomy and sub-category, then that record must contain characters

5   which are responsive to the user's search. After a match is detected, computer 10 compares the

search term 3001 against only those records having matching taxonomies/categories.

Once the matching records have been identified, computer 10 generates a numerical count

of all of the records of biological data within bioinformatics data collection 1 which have

characters which match the search term. This numerical count is further broken down by sub-

10   category. For example, Figure 4 shows "3,501" unique genetic records for the category "Cell

Growth and Maintenance" 502. Within this, "105" relate to sub-category "Oncogenesis" 504.

In another embodiment of the invention, computer 10 launches a search query comprising

only a category or sub-category without a search term. This enables a user to "drill-down"

through bioinformatics data collection 1 merely by selecting a narrower and narrower sub-

15   category. In yet another embodiment of the invention, computer 10 is adapted to launch search

queries comprising only a search term or terms. It should be noted that computer 10 initiates any

one of these types of search queries at any level of drill-down.

In an illustrative embodiment of the present invention, a user may also drill-up through a

hierarchy of categories/sub-categories. For example, once a user has drilled down and reached the

20   level represented by screen 4000b in Figure 4, he/she may click on the category "Biological

Process" 505, and upon receiving this category as search-related data, computer 10 returns to

screen 4000a in Figure 4. In addition to drilling-up, the user 3 may switch taxonomies at any

point in a drill-down or up. For example, the user can click on the taxonomy "Molecular

21

Function" 4002 or "Cellular Component" 4003 in Figure 4 and be presented with categories

corresponding to this taxonomy and all previous search constraints are maintained. In all cases,

when the user clicks on or otherwise selects a taxonomy, category or sub-category, computer 10

compares the search-related data to a hierarchy as previously explained. A search is then

5   launched by computer 10 using navigational sub-categories which result from this comparison.

Figures 5 and 6 display screens 5000 and 6000 depicting other examples of how results

from a search using two or more taxonomies 5001, 5002 and 5003 can be displayed. Beginning

with Figure 5, there is shown an example of an initial screen 5000 which displays categories 505

which make up a "Biological Process" taxonomy 5002. Though only a few categories are shown,

10   it should be understood that categories 505 may comprise any topic, or some subset. In the

example shown in Figure 5, the user types in a search term "acid" 3002 and then clicks on the

"Molecular Function" taxonomy 5001. The present invention, however, is not limited to

displaying the results of a search against only one taxonomy on one screen at the same time.

Rather, the present invention can display the results of searches against multiple taxonomies on

15   one screen at the same time.

Computer 10 then selects navigational sub-categories 506 which correspond to the

taxonomy "Molecular Function" and subsequently launches a search query against bioinformatics

data collection 1 using search term 3002, taxonomy 5001 and sub-categories 506. It should be

noted that all three taxonomies 5001, 5002 and 5003 and are provided to enable a user to initiate a

20   search using any taxonomy.

Continuing, Figure 6 depicts an example of a screen 6000 generated from the results of

initiating the just described search query. As shown, the screen 6000 displays categories 506

which are navigational sub-categories related to the taxonomy "Molecular Function" 5001. In

22

addition, the number of records containing characters matching the search term "acid" 3002 is also displayed. As before, this number is displayed as a total and is also broken down for each sub-category. For example, next to the sub-category "Structural Protein" 5004 is the number "12" which indicates the number of genetic records with a molecular function of structural protein that contain the character-string "acid" and are contained in bioinformatics data collection 1.

It should be understood that the user need not input an additional keyword to further narrow his/her search. Instead, computer 10 generates intuitive sub-categories 506 which are presented to the user for the very purpose of narrowing his/her search. In addition, the number of matching records for each sub-category is displayed without the need for the user to individually launch separate searches aimed at each sub-category.

It should be understood that the terms "category" and "sub-category" are relative terms and in some instances may be used interchangeably.

The ability to switch among taxonomies, to drill-down or up, or to switch among taxonomies while drilling down or up enables the user to navigate a Web site or other user interfaces and corresponding bioinformatics data collection 1 with great ease. This ease-of-navigation can be used to enable new revenue models. In one embodiment of the invention, new revenue models, such as advertising models, are enabled from such easy-to-navigate Web sites.

Figure 7 provides a schematic of the data as it is stored and organized in a bioinformatics data collection in accordance with a preferred embodiment of the present invention. The bioinformatics data collection 705 contains many records of biological data, 705a, 705b, and 705c. In this example, a record is a single unit of identifiable data.

Three exemplary records are shown in Figure 7. Each of records 705a, 705b and 705c is a particular gene available in the bioinformatics data collection.

Indices 710, 715a and 715b are used to access records in bioinformatics data collection 705. Inverted index 702 contains a listing of all the key words and phrases 710 in all of the records of biological data in bioinformatics data collection 705, and other indices 715a and 715b. Examples of such key words and phrases include "cytoplasm," "karogamy," "peptidylprolyl," "zygote," "adenylate" and "7SLRNA." Attached to each of these key words and phrases are links 710b. These links reference each record in index 705 that contains these character-strings.

Indices 715a and 715b represent different taxonomies of bioinformatics data collection 705. As shown by the headings, index 715a is a "Biological Process" taxonomy of bioinformatics data collection 705 and index 715b is a "Molecular Function" taxonomy of bioinformatics data collection 705.

These three indices 710, 715a and 715b are used to access the records in bioinformatics data collection 705 in three different ways. Index 710 receives search terms or phrases and is scanned to locate those key word or phrases. When a hit is discovered, the number of links 710b that reference into bioinformatics data collection 705 is then determined.

Indices 715a and 715b provide record collection lists of their respective contents in response to user input. As an example, if the user clicks on the "Biological Process" taxonomy, all of the categories within that taxonomy are displayed. Two of those categories include "Cell Growth & Maintenance" and "Cell Communication." As shown in Figure 7, each of these categories is divided into sub-categories like "Meiosis," "Membrane Fusion," "Metabolism," "Cell Recognition," "Cell-Cell Signalling" and "Signal Transduction."

Index 715b is a taxonomy of bioinformatics data collection 705 based on "Molecular Function." Within taxonomy 715b are categories. The exemplary categories are price ranges by dollar amount.

By having multiple taxonomies of the single database, multiple paths are possible to reach the same records. Figure 10 shows one set of queries from a user and the system responses that represent a path a user may take to reach the records he/she desires. The user begins by typing in a search term against the "Biological Process" taxonomy. In the example given the search term is

5   "acid." The present invention queries term index 710 and determines that 2,007 records in the database have the word "acid" within them.

The present invention then determines the categories that are associated with the search term "acid". For example, almost all of the records that have the search term "acid" in them are categorized into the group of "Cell Growth and Maintenance." The user selects the "Cell Growth

10   and Maintenance" category and the present invention then searches through index 715a to determine how many records within each of the sub-categories also are associated with the search term "acid." As shown in Figure 8, only 8 records organized into the "Budding" category contain the keyword "acid" while 38 records organized into the "Cell Cycle" category contain the keyword "acid." Thus the present invention compounds all of this data and provides it to the user.

15   It should be noted that by pushing data back to the user, in this case a glimpse of the organization of the categories, the user can learn how best to proceed with drilling down into the data.

The user responds to the list of sub-categories provided by the present invention by selecting one. In this example, the user selects the sub-category "Cell Cycle".

The system responds by providing a list of all 38 records that are associated with the

20   search term "acid." To narrow the list further, the user clicks on the "Molecular Function" taxonomy in response.

The system responds by cross-matching the 38 records against the categories within the "Molecular Function" taxonomy. Thus, the system generates a data collection of these 38 records as organized by molecular function (i.e., enzyme has 14, etc.).

The user responds to these sub-categories by selecting a particular molecular function, say "Nucleic Acid Binding". The system responds by cross-matching the sub-categories within "Nucleic Acid Binding". Once the cross-matching is completed, the system provides the user with a list of appropriate sub-categories with how many records match the search so far.

The user responds by selecting "RNA Binding". The system responds by providing a list of the one record that matches the search. Thus, the listed records are a match of the taxonomy "Biological Process;" the search term "acid;" the category "Cell Growth & Maintenance;" the sub-category "Cell Cycle;" the taxonomy "Molecular Function;" the category "Nucleic Acid Binding;" and the sub-category "RNA Binding."

Figure 11 shows another set of user queries and system responses that represent another path the user may use to get to the same set of records. The user begins this search by requesting details about the "Molecular Function" taxonomy. The system responds by returning the list of molecular functions with a count of how many records are associated with each function.

The user responds by entering the search term "acid." The system cross-matches the search term "acid" in free-text term index 710 with each molecular function. This produces a category list of molecular functions with the number of records associated with the search term "acid" in parentheses.

The user responds by selecting one of the listed categories. Following with the example given in conjunction with Figure 10, the user selects "Nucleic Acid Binding."

The system responds by providing a list of sub-categories under the category "Nucleic Acid Binding." The user responds by selecting a sub-category, such as "RNA Binding."

The system responds by providing a list of all 120 genetic records under "RNA Binding" that are associated with the search term "acid." The user responds by selecting the "Biological Process" taxonomy. The system responds by cross-matching all of the categories in the "Biological Process" taxonomy with the selected sub-category "RNA Binding." Thus, the system generates a data collection of these 120 records as organized by biological process (i.e., "Cell Growth & Maintenance has 35, etc.).

The user responds to these categories by selecting "Cell Growth & Maintenance." The system responds by cross-matching the sub-categories within "Cell Growth & Maintenance." Once the cross-matching is completed, the system provides the user with a list of appropriate sub-categories with how many records match the search so far.

The user responds by selecting "Cell Cycle." The system responds by listing the one record that matches that search. In this example, the records match the taxonomy "Molecular Function;" the search term "acid;" the category "Nucleic Acid Binding;" the sub-category "RNA Binding;" the taxonomy "Biological Process;" the category "Cell Growth & Maintenance;" and the subcategory "Cell Cycle." This is a different search path to the one described in Figure 10, yet it yields the same result.

Figure 12 shows yet another set of user queries and system responses that represent yet another path the user may travel in order to obtain the desired records. The user begins by selecting the "Biological Process" taxonomy. The system responds by listing all of the categories with all the records associated with each category in parentheses. In this example, each biological process category is listed along with its number of associated records.

27

The user responds by selecting one of the listed categories. Again, the user selects "Cell Growth & Maintenance." The system responds by listing the sub-categories under the selected category along with the number of associated records in parentheses.

The user responds by entering the search term "acid." The user responds by entering the search term "acid." The system cross-matches the search term "acid" in free-text term index 710 with each sub-category under "Cell Growth & Maintenance." This produces a list of sub-categories under "Cell Growth & Maintenance" with the number of records associated with the search term "acid" in parentheses.

The user responds by selecting the "Molecular Function" taxonomy. The system responds by cross-matching all of the categories in the "Molecular Function" taxonomy with the records associated with the search term "acid" that are contained in the category "Cell Growth & Maintenance" The system then provides the user with a list of categories in the "Molecular Function" taxonomy. Examples of categories in this taxonomy are "Ligand Binding or Carrier", "Motor" and "Nucleic Acid Binding."

The user responds by selecting a particular category. Following with the above examples, the user selects the category "Nucleic Acid Binding." The system responds by providing the sub-categories within the category "Nucleic Acid Binding." The number in the parentheses corresponds to the number of records that are associated with the category "Cell Growth & Maintenance" and each of the listed sub-categories within this category of "Nucleic Acid Binding" that contain records associated with the search term "acid" (i.e., "DNA Binding," "Ribonucleoprotein," and "RNA Binding").

The user responds by selecting the sub-category "RNA Binding." The system responds by providing a list of all of the records that match the search. The user refines the search via the

"Biological Process" taxonomy. Thus, the user selects the "Biological Process" taxonomy and the system responds by cross-matching the records associated with the sub-category "RNA Binding" with the categories of the "Biological Process" taxonomy. The system then displays the listing of categories with the number of records associated with the sub-category "RNA Binding" and each biological process under the "Cell Growth & Maintenance" category that are associated with the search term "acid."

Thus, the system responds by listing the sub-categories under the category "Cell Growth & Maintenance" (i.e., "Cell Cycle," "Meiosis," "Metabolism, " etc.) with the number of records associated with "RNA Binding" in parentheses.

The user selects a listed sub-category. Following the above example, the user selects "Cell Cycle." The system responds by listing all of the "RNA Binding" associated records that are also associated with "Cell Cycle" and the search term "acid." This yields the one result that matches the search. In this example, the listed records match the taxonomy "Biological Process;" the category "Cell Growth & Maintenance;" the search term "acid;" the taxonomy "Molecular Function;" the category "Nucleic Acid Binding;" the sub-category "RNA Binding;" the taxonomy "Biological Process;" and the sub-category "Cell Cycle." This is a different search path to the one described in Figures 8 and 9, yet it yields the same results.

These three examples demonstrate the versatility of the present invention. First, the user is not required to go through a specific path to reach the desired number of records. While the above examples show only three paths to reach the desired set of records, it can be appreciated that there are multiple paths to reaching the same set of records.

This plurality of paths is achieved by the independence of the taxonomies shown in Figure 7. By keeping these taxonomies independent, the user may switch among which taxonomy he/she

29

wishes to use to consider the data and make queries into bioinformatics data collection 705. The level of the search that the user uses to make a decision to switch among available taxonomies is also arbitrary and up to the user. This allows users who are more proficient in developing searches to use their proficiency in one taxonomy index to whittle the number of records down

5   before going into another taxonomy index to finish the search where the user is less proficient, and vice versa.

Another feature of the present invention is the pushing of data to the user. As noted above, the user receives category and sub-category information when a query via a search term is used earlier in the process. As noted above, suppose the user is looking for purine metabolism,

10   instead of acid. By typing the search term "puring," the system will provide the category list to the user so that he/she can drill down into the data. Thus, if there were a sub-sub-category of "metabolism" the user would eventually see that sub-sub-category and make the association between "purine" and "metabolism." Thus the user comes in contact with a useful category or sub-category that he/she can use to search for desired information. Additionally, if the character-

15   string "purine" were contained in any genetic record description, such genetic record would appear in the search set following the user's entry of such keyword query.

These records are categorized so that associations are made between the categories and sub-categories in the multiple taxonomies and the records. In addition, terms within the records that correspond to terms in the free text term index are determined. Associations are then made

20   between these records and the various categories and terms in the indices.

Another advantage of the present invention is the way results are provided to the user. As noted in the many examples above, much of the sifting through the bioinformatics data collection is done via the categories and sub-categories. In a preferred embodiment, there are many more

30

records in the bioinformatics data collection than there are categories. As an example, a search term may be associated with thousands of records, but only one category. Providing a list of thousands of records requires a lot of data handling in both the transmission of the data to the user, as well as the displaying of the data to the user. Providing a list of only one category is much less data to transmit and display. This makes the invention ideal for use with devices with small screens, such as cell phones, pagers, and personal digital assistants (PDAs) and palm-held devices.

Figure 14 is a representation of a portion of the data stored in structure 702 and how that data is organized in accordance with a preferred embodiment of the present invention. Node 1405 represents the category "Cell Growth & Maintenance" from the "Biological Process" taxonomy. Node 1410 represents the sub-category "Metabolism." Node 1415 represents the sub-category "Cell Cycle." Node 1420 represents the sub-category "Enzyme" from the "Molecular Function" taxonomy. Record 1425 represents a single record.

Linking the nodes and records are category code words. Leading into node 1405 is a category code word called "CG." Leading into node 1410 is a category code word called "ME." Leading into node 1415 is category code word "CC." Leading into Record 1425 are links R1 and R2. This representation shows how the various categories relate to each other and the records.

In one embodiment of the present invention, these path names are stored in inverted index 702 and used to retrieve records. This structure provides several advantages. In one embodiment of the present invention, these path names are stored in inverted index 702 and used to retrieve records. This structure provides a means to perform Boolean operations on the path names to calculate category count results and to identify records that are identified by those category paths.

It will be appreciated that large global collections of data can be broken down into smaller sub-collections. The sub-collections can be stored independently one from the other, as in separate physical locations or simply in separate data tables within the same physical location, and can be connected one to the other through a network. As data are added to the large global collection overall, it can be sent and added to individual sub-collections and/or can be formed into a further sub-collection. For instance, data entered by educational institutions and scientific research facilities can be stored independently in their own data storage facilities and connected to one another via a network, such as the Internet. Thus, as can be seen, the present invention can be implemented with very little or no change in the present protocol for data collection and storage.

It will be appreciated that the present invention provides a search interface that can aggregate disparate databases and make the disparate databases searchable through one interface.

Once the individual sub-collections have been identified, each performs its own indexing function. In carrying out the indexing function, each sub-collection creates its own sub-collection taxonomy consisting of statistical information generated from what is commonly referred to as an inverted index. An inverted index is an index by individual words listing records which contain each individual word. The indexing function itself can be carried out in any method. For example, indexing can be performed by assigning a weight to each word contained in a document. From the weights assigned to the words in each document, a sub-collection view (i.e., the statistical information derived from the inverted index) is created upon completion of the indexing function. Regardless of how the sub-collection indexing is carried out, each sub-collection will have its own independent sub-collection view based upon that sub-collection's inverted index. When data information is added to the sub-collection, the indexing function is carried out again and the sub-collection's view can be re-compiled from a new inverted index.

32

Upon completion of each sub-collection view, certain statistical information about the sub-collection view is gathered by a global collection manager to form a global collection of parameters, statistics, or information. The global collection manager may either request from each sub-collection that it send its sub-collection view or certain statistical information about the sub-collection, and/or each of the sub-collections may spontaneously send the sub-collection view or certain statistical information about the sub-collection to the global collection manager upon completion. Regardless of whether the taxonomies are requested or spontaneously sent, upon collection at the global collection manager of all of the sub-collection's views or certain statistical information about the sub-collection, the global collection manager builds a "global view" or certain statistical information about the global view on the basis of the sub-collection views or certain statistical information about the sub-collection. Necessarily, the global view is likely to be different from each of the individual sub-collection views. Once the global view or certain statistical information about the global view has been compiled, it is sent back to each of the sub-collections. Figure 20 represents the global view. This is not the case in this current embodiment although it could be the case in another embodiment.

In this manner then, a distributed data retrieval system is built and is ready for search and retrieval operations. To search for a particular piece of data information, a system user simply enters a search query. The search query is passed to each individual sub-collection and used by each individual sub-collection to perform a search function. In performing the search function, each sub-collection uses the global view to determine search results. In this manner then, search results across each of the sub-collections will be based upon the same search criteria (i.e., the global view).

33

The results of the search function are passed by each individual sub-collection to the global collection manager, or the computer which initiated the search, and merged into a final global search result. The final global search result can then be presented to the system user as a complete search of all data information references.

These time savings are increased as the length of the path is increased. If the entire path length from base node to document node includes fifty of these node-to-node or node-to-document links, the search is reduced from 400 characters to 100.

The labeling of these paths also reduces computation time for other searches. For example, if the search is a proximity search (i.e., Is gene X trans to gene Y?), the present invention can be used to make this determination. .

It should be noted that other variations are possible with this embodiment of the invention without departing from the scope of the invention. For example, the number of characters used to describe a path is not limited to two and may in fact be any number of characters. Additionally, the path names need not be limited to letters but may encompass numbers, symbols or a combination of letters, numbers and symbols. In addition, once the paths between the base node and each document are determined, they may be stored within the records as tags in a preferred embodiment of the present invention.

Figure 11 shows a system overview in accordance with an embodiment of the present invention. Hub computer 505 is the central point. It receives queries from and provides compiled results to users. Hub computer 505 is comprised of front end 505a, back end 505b, microprocessor 505c and cache memory 505d. Front end 505a is used to receive queries from users and format the results so that they are in a compatible format for the user to understand. Back end 505b uses the appropriate protocols to issue broadcast messages and receive messages.

34

Coupled to hub computer 505 are spoke computers 510a, 510b through 501n. Spoke computers 510a-510n have local memories 510a1-510n1 that are used to store indices. Coupled to each spoke computer 510a-510n is large memory storage 515a-515n used to store the records in bioinformatics data collection 705.

5   In a preferred embodiment of the present invention, hub computer 505 and spoke computers 510a-510n are Intel-based machines. The communications between the hub computer 505 and spoke computers 510a-510n are based on the TCP/IP format. Spoke computers 510a-510n operate using a custom software written in C++ or Visual Basic. Hub computer 505 uses Visual Basic and C++ to process data.

10   Figures 15 through 20 show a method and an apparatus for the efficient and effective distribution, storage, indexing and retrieval of data information in a distributed data retrieval system which is fault tolerant. Large amounts of data may be searched faster by distribution of the data, separate indexing of that distributed data, and creation of a global index on the basis of the separate indexes. A method and apparatus for accomplishing efficient and effective distributed

15   information management will thus be shown below.

Referring to Figures 15 and 16, in step 100 of Figure 15 data information is distributed and formulated into sub-collections 150 of Figure 16. The process of distributing the data may be accomplished by sending the data from a central computer terminus 110 to local nodes 120, 130 and 140 of a computer network 10, or by directly entering the data at the local nodes 120, 130 and

20   140. Further, the data may be divided such that the divided data is of equal or unequal sizes, and so that each division of the data has a relational basis within that division (i.e., each division having an informational subject relation all its own). Such allowances for data entry and distribution allow for little or no change to current data entry and distribution protocols. In the

case of the Web, data entry can continue as it does now. Each entity (i.e., Manufacturers,

Distributors, Retailers, etc.) can continue to enter data as it sees fit. Thus, the sub-collections 150

can be organized in any fashion and be of any size.

In step 200 of Figure 15, the data information, which has been divided and stored into the

sub-collections 150, is indexed and a "sub-collection view" is formed. Indexing of the sub-

collection 150, like the step of distributing the data, can follow current protocols and may be

computer-assisted or manually accomplished. It is to be understood, of course, that the present

invention is not to be limited to a particular indexing technique or type of technique. For instance,

the data may be subjected to a process of "tokenization". That is, records containing the data are

broken down into their constituent words. The resulting collection of words of each document is

then subject to "stop-word removal", the removal of all function words such as "the", "of" and

"an", as they are deemed useless for document retrieval. The remaining words are then subject to

the process of "stemming". That is, various morphological forms of a word are condensed, or

stemmed, to their root form (also called a "stem"). For example, all of the words "running", "run",

"runner", "runs", . . . , etc., are stemmed to their base form run. Once all of the words in the

document have been stemmed, each word can be assigned a numeric importance, or "weight". If a

word occurs many times in the document, it is given a high importance. But if a document is long,

all of its words get low importance. The culmination of the above steps of indexing convert a

document into a list of weighted words or stems. These lists of weighted words or stems are thus

in the form:

document.sub.1 .fwdarw.word.sub.1, weight.sub.1 ; word.sub.2, weight.sub.2 ; . . . ;

word.sub.n, weight.sub.n.

Alternatively, the same indexing of the sub-collection can also be achieved using a bit-mapped indexing technique.

Regardless of the indexing technique used above, the index thus far created is then inverted and stored as an "inverted index", as shown in Figure 15. Inversion of the index requires pulling each word or stem out of each of the records of the index and creating an index based on the frequency of appearance of the words or stems in those records. A weight is then assigned to each document on the basis of this frequency. Thus, the inverted index, has the form of:

word.sub.1 .fwdarw.document.sub.a, weight.sub.a ; document.sub.b, weight.sub.b ; . . . ; document.sub.z, weight.sub.z.

The inverted index 210 itself, as shown in Figure 15, is composed of many inverted word indexes 220, 230 and 240, and can thus be created and organized. As shown, each inverted word index 220, 230 and 240 composes an index of a different word, taken from the records of the initial index, such that each document is weighted in accordance with the frequency of appearance of the word in that document. Completion of the inverted index 210 allows the derivation of statistical information relating to each word and thus the creation of a sub-collection view 410, as shown in Figure 20. The statistical information which makes up the sub-collection view 410 includes the total number of records in the sub-collection 150 and, relating to each word, the number of records in the sub-collection that contain that word. As each computer is indexing its sub-collection separately, the total indexing time for indexing the entire collection is greatly reduced as it is now shared across many computers. It is to be understood, of course, that any method of indexing may be used to form the sub-collection view 410 and that the above described method is but one of many for accomplishing that goal.

In step 300 in Figure 15, once the sub-collection view 410 is created, a global view is created and distributed. For formation of the global view, each sub-collection view 410 which has been created is collected from the local nodes 120, 130 and 140 of the computer network 10 and sent to the central computer 110. Referring to Figure 19, showing an embodiment of the paths of

5   communication of a computer network 20, sub-collection views from computers 320, 330 and 340 are sent to central computer 310 along communication paths 4.1. Collection and sending of the sub-collection view can be initiated by either the central computer 310 or the local computers 320, 330 and 340. If collection of the sub-collection views 410 is initiated by the central computer 310, it may be initiated by individual commands sent to each computer in the network 20, or as a group

10   command sent to all of the computers in the network 20. If the collection of the sub-collection views 410 is initiated by the local computer 320, 330 or 340, then the local computer may send the sub-collection view upon occurrence of completion of the sub-collection view, an update of the sub-collection view, or some other criteria, such as a specific time period having elapsed, etc. It is to be understood, of course, that any method by which the completed sub-collection views are

15   sent to the central computer from the local computers is acceptable.

Upon collection of all of the sub-collection views 410, a global view 510 is created as shown in Figure 20. In the formation of the global view 510, the central computer 310 uses the sub-collections 410 that have been sent from every local computer 320, 330 and 340 to determine how many records are contained in the sub-collection residing at the particular local computer,

20   and for every word, how many records in the sub-collection contain the word in question. The global view 510 then comprises information pertaining to how many records there are in all of the sub-collections (i.e., the total document sum) and for every word, how many records in all of the sub-collections contain the word in question. The global view, then, provides all of the necessary

38

information for use in weighting the words in a user query, as will be explained below. It is to be understood, of course, that any method which provides the central computer with the information necessary to form the global view may be used. For instance, the sub-collection views need not be sent in their entirety themselves, but instead the nodes could send only statistical information about their subcollection(s).

To complete step 300 of Figure 15, the global view 510 is sent from the central computer 310 to each of the local computers 320, 330 and 340 by way of communication paths 4.2 (as shown in Figure 21). Thus each local node in the network will now have the global view. It is to be understood, of course, that the description of the formation of the sub-collection views and subsequent formation of the global view can be conducted on any computer network, and thus computer networks 10 and 20 are to be considered interchangeable in this description.

In step 400 of Figure 15, the search phase is conducted. The search phase refers to search and retrieval of data information stored in the large data text corpora. Thus, to begin with, in the search phase a search query is entered and uploaded by a system user into the computer network 10. It is to be understood, of course, that the system user may enter the search query at any computer location that is connected to the computer network 10. Upon entry of the search query, the search query is transmitted by the computer network 10 to all of the local computers 120, 130 and 140 in the computer network 10.

After receiving the search query, each local computer 120, 130 and 140 then indexes the search query using the same steps that are used to index the records, namely, for instance, "tokenization", "stop word removal" and "stemming" and "weighting". The resulting words (actually stems) in the query are assigned importance weights using the global view 510 which each local computer 120, 130 and 140 received in step 300. If a query word is used in many

39

records, then it is presumed to be common and is assigned a low importance weight. However, if a handful of records use a query word, it is considered uncommon and is assigned a high importance weight. The "total number of records in the collection" and the "number of records that use the given word" statistics are only available to local computers 120, 130 and 140 after the global view creation.

It is to be noted, of course, that other formulae might be used as desired. If so, the sub-collection view may be adjusted to account for the different formula. It should also be noted that having each local computer perform an indexing of the search query might be necessary if the entry point of the search query is at a point which does not have access to the global view and thus cannot perform the indexing function. However, if the entry point for the search query does have access to the global view, then the search query can be indexed at the entry point and distributed in an indexed format.

The indexing of the search query, as shown above, yields a weighted vector for the search query of the form:

query.fwdarw.word.sub.1, weight.sub.1 ; word.sub.2, weight.sub.2 ; . . . ; word.sub.n, weight.sub.n.

Having indexed the search query, a simple formula is used to assign a numeric score to every document retrieved in response to the search query. This simple formula, referred to as a "vector inner-record similarity" formula can assign a weight to a word in the search query and another weight to a word in the document being scored. Each document is then sent to the central computer 310, via communication paths 4.1, from the local computer nodes 320, 330 and 340.

In step 500 of Figure 15, once all search results have been returned to the central computer via communication paths 4.1, the central computer 310 merges the variously retrieved records into

40

a list by comparing the numeric scores for each of the records. The scores can simply be compared one against the other and merged into a single list of retrieved records because each of the local computers 320, 330 and 340 used the same global view 510 for their search process. Upon completion of the merging of the records, a complete list is presented to the system user. How

5   many of the records are returned to the user can, of course, be pre-set according to user or system criteria. In this manner then, only the records most likely to be useful, determined as a result of the system user's search query entered, are presented to the system user.

It should be noted that the manner in which the global view 510 is created provides a fault tolerant method of distributing, indexing and retrieving of data information in the distributed data

10   retrieval system. That is, in the case where one or more of the sub-collection views is unable to be collected by the central computer, for whatever reason, a search and retrieval operation can still be conducted by the user. Only a small portion of the entire collection is not searched and retrieved. This is because failure by one or more local computers results in only the loss of the sub-collections associated with those computers. The rest of the data text corpora collection is still

15   searchable as it resides on different computers.

Further, to provide even more fault tolerance, data information may be duplicatively stored in more than one sub-collection. Duplicative storage of the data information will protect against not including that data information in a search and retrieval operation if one of the sub-collections in which the data information is stored is unable to participate in the search and retrieval.

20   Thus the foregoing embodiment of the method and apparatus show that efficient and effective management of distributed information can be accomplished. The current invention of the division of the large data text corpora into sub-collections which are then separately indexed, which indexes are then used to form a global view, is possible, as shown herein, without a loss

41

and, in fact, an increase in the effectiveness and efficiency of a search and retrieve system. Further, the search and retrieval operations take less time than current systems which either search the entire large collection all at once or which search individual collections.

This system implements the search queries described above in the following manner.

5 First, hub computer 505 receives a query from the user. This query can be in the form of a search term, a taxonomy selection, a category selection, a sub-category selection, etc. Upon reception of the query, microprocessor 505c compares the query with data stored in cache 505d. If the response to the query is already stored in cache 505d, the microprocessor 505c returns that response as a result to the user. Hub computer 505 then waits for another query from the user.

10 If the query is not in cache 505d, microprocessor generates a broadcast message to be sent to all spoke computers 510a-510n. This broadcast message includes the user's query.

Upon reception, each spoke computer 510a-510n performs a search of the appropriate index stored therein using the query from the user. In a preferred embodiment of the present invention, each spoke computer 510a-510n stores all three indices 710, 715a and 715b in local memory as described above. In addition to broadcasting a request across the network to different machines, multiple threads could be used and the message could be broadcast to multiple processors in a single machine (on a bus rather than a network). Alternatively, the search request could be conducted locally -- a single process, single thread, single machine search.

Also in the preferred embodiment, data storage 515a-515n each stores only a portion of the records in bioinformatics data collection 705. Since each set of data is unique in data storage 515a-515n, it follows that the relationships between the indices stored in local memories 510a1-510n1 are also unique because they cannot all access the same records. In an alternate embodiment, spoke computers 515a-515n all share identical copies of bioinformatics data

42

collection 705, but the indices 710, 715a, and 715b are parsed among local memory 510a-510n

Each spoke computer 510a-510n returns the results, either a list or the counts for each category, determined by its respective indices to hub computer 505. Hub computer 505 compiles those results and provides them to the user. In an alternate embodiment, spoke computers 515a-

5   515n are also provided with cache memories to reduce the number of queries made to memories 515a-515n.

Figure 14 is a system in accordance with the present invention. At block B1405, the system receives a query from the user. It should be noted that the query may be a term, a taxonomy, a category, a sub-category, a sub-sub-category, free text, a field, a numeric range,

10   Boolean logic, combinations of elements, etc. At block B1410, the query is formulated with respect to the current state of the present search. As an example, if the user enters a keyword query, the query is formulated such that the current taxonomy is taken into consideration.

At block B1415, the system determines the appropriate categories or sub-categories to search through to locate records that match. As an example, one possible category is "Pants."

15   From the determinations made in blocks B1410 and B1415, the system has narrowed the number of possible hits by discarding those records that do not conform to the selected category. It should be noted that, in a preferred embodiment, the categories or sub-categories are determined using an organized list such as a B-tree, another bioinformatics data collection or from the inverted index itself.

20   At block B1420, the system checks its cache. The cache typically stores three types of data. The first type of data is a query result that was recently performed. Thus if user A issues a query for term X in category Y, and 1 minute later user B makes the identical query, the cache is used to provide the results, instead of determining the results anew. The second type of data

43

stored in the cache is frequently requested queries. Suppose users are, in the aggregate, frequently requesting records on new cars but not requesting records on the disease malaria. The results from this frequently requested query are then stored in the cache. The third type of data is searches that are precompiled because otherwise they would take a long time to perform.

5   If the query is not in the cache, then the query is broadcast to a plurality of processors operating in parallel at block B1425. It should be noted that blocks B1425, B1430 and B1435 are in dashed lines because they are not requirements of the process in order to be operational, but rather are preferred embodiments that enhance the performance of the process. To be more specific, if the query is found in the cache, then blocks B1425-B1435 are eliminated and the overall time to provide the user with results is reduced. The use of parallel processors operating on either portions of the query or searching only portions of the inverted index also reduces the amount of time it takes to provide a result. Thus, a slower performing system that did not include a cache or parallel processors could also use the present process to generate results.

        At block B1430, the system receives the number of records that "hit" on the query provided in block B1405. At block B1435, the hits are compiled and the number of hits per category, as determined in block B1415, is also compiled.

        At block B1440, the results are displayed to the user. Typically, these results are organized into categories. However, in a preferred embodiment, the system will display a default list of document hits when there are no sub-categories below the last category selected by the user. This prevents giving the user a listing of categories with 0 document hits because this information is not as useful to the user as to know which category the document hits are located in.

        At block B1445, a determination is made based upon the results displayed. If the user is satisfied with the results, the process ends at block B1450. If the user desires to refine the query

44

or drill-down or drill-up further into the bioinformatics data collection, the process continues with a new query at block B1405.

Figure 13 is a screen shot of a categorizer in accordance with an embodiment of the present invention. This embodiment of a categorizer is a graphic user interface (GUI) that a system operator uses to assist in associating records with categories. Typically, the system operator uses this embodiment of the present invention to insert a new document into an existing category in the taxonomy. Section 1305 is a toolbar that provides such functionality as editing, searching within a document, changing the viewed document, printing, etc. Section 1310 is a graphic representation of the categories in the taxonomy. Section 1315 is a display of the current document.

The system operator scrolls through the taxonomy in section 1310 and the document in section 1315 looking for the best-fit categories for the document displayed in section 1315. When the system operator believes he/she has found a best-fit category for the displayed document, he/she instructs the system to make an association between the best-fit category and the displayed document by clicking button 1320.

In a preferred embodiment of the present invention, the document is scanned by the system before it is displayed. This scanning procedure compares the key terms stored in 710 with the word in the document. When a match is made, the document is highlighted so that the system operator may quickly discern which key terms are in that document. In addition, a count is performed on how many key terms are in this document. The system then queries the various category indices looking for a category title that matches the key term with the most hits in the document. Once that category is determined, that category is displayed along with its parent categories and its sub-categories so as to provide a frame of reference for the system operator. If

the system operator agrees with the automatically determined category, he/she clicks on button 1320 to create an association between that determined category and the displayed document. If the system operator does not agree with suggested category and cannot find another suitable category by searching through the list of categories, he/she clicks on button 1325 to instruct the system to create a new category into the hierarchy.

The present invention is not limited to those embodiments described above. For example, the search terms entered by the user need not only be textual. The present invention also includes embodiments that can perform searches on number ranges, proximity, field searches and Boolean searches. In addition, the present invention may be used with other types of queries such as natural language and context-sensitive queries.

Another embodiment of the present invention includes alternative queries placed into the cache. For example, before the first query is processed, precompiled queries such as those that are known to take a long time or are particularly timely, can be pre-loaded into the cache to save time.

The present invention is also not limited to three taxonomies. Any bioinformatics data collection can be represented by an unlimited number of taxonomies. Alternative embodiments are envisioned that include viewing records by other identifiable category structure. Moreover, there is no theoretical limit to the depth of sub-categorization for each taxonomy.

The present invention is also not limited to when certain taxonomies are provided to the user. As described above, the user is presented with the taxonomy last selected. Thus, if the user is using the "Biological Process" taxonomy and enters a new search term, the results will be displayed following the "Biological Process" taxonomy described above. However, in an alternative embodiment, the system can switch taxonomies automatically for the user in an effort to present the search results in a more meaningful manner. For example, if the user selects the

46

final sub-category in the chain, the system will automatically switch over to another taxonomy so as to provide the user with more context and scope regarding the remaining search results. Thus, if there are no sub-categories under a "Biological Process" category the present invention will switch the taxonomy to a different taxonomy so that the user is provided with greater context and scope regarding the remaining search results. This switching can also be based on the number of hits. If the category contains only two hits, the system will automatically switch to a different taxonomy to provide the user with more useful information on the remaining records. Similarly, the automatic taxonomy switching may also be based on a particular taxonomy where the number of categories or sub-categories is small. For instance, providing the user with the information that all the hit records are located in one category does not provide any information the user can use to distinguish between these records. Switching to another taxonomy may provide the user with more categories he/she can use to distinguish between the hit records.

It will be appreciated that there is no limit to the depth of the categories and sub-categories. Additionally, it will be appreciated that the present invention can be implemented in an interface other than the Web.

It will further be appreciated that one preferred embodiment of the present invention is a system for searching a collection of data, said system comprising: an organizer configured to receive search requests, said organizer comprising: a collection of data having at least two entries; wherein the collection of data is organized into at least two taxonomies; wherein each of the at least two taxonomies is associated with at least two categories; wherein the entries correspond to at least one of the at least two taxonomies and also correspond to at least one of the at least two categories; and a search engine in communication with the collection of data, wherein said search engine is configured to search based on the at least two taxonomies and based on the at least two

47

categories, wherein the search engine returns, in response to a search request identifying at least a first taxonomy of the at least two taxonomies, a list of the categories associated with the at least first identified taxonomy, along with the number of entries associated with each of the categories associated with the at least first identified taxonomy.

5   In a preferred embodiment of the present invention, the returned list of categories associated with the first taxonomy, along with the number of entries associated with each of the categories associated with the identified taxonomy can be further searched with regard to a second of the at least two taxonomies, whereby the search engine returns, in response to a search request identifying the second taxonomy of the at least two taxonomies, a list of the categories associated

10   with all identified taxonomies, along with the number of entries associated with each of the categories associated with the second taxonomy.

In another preferred embodiment, the search engine, having returned, in response to a search request identifying a first taxonomy of the at least two taxonomies, a list of the categories associated with the identified taxonomies, along with the number of entries associated with each

15   of the categories associated with the identified taxonomies, will provide only those categories with a non-zero number of entries associated with the identified taxonomy and will further return sub-categories both associated with the category and having a non-zero number of entries associated with the sub-category.

Still further in another preferred embodiment, the search engine, having further returned

20   sub-categories both associated with the category and having a non-zero number of entries associated with the sub-category, will, in response to a search request identifying a second taxonomy of the at least two taxonomies, provide a list of the categories with a non-zero number

48

of entries associated with the at least second identified taxonomy, along with the number of entries associated with each of the categories associated with the second identified taxonomy.

In another embodiment, the search engine, having returned, in response to a search request identifying a first taxonomy of the at least two taxonomies, a list of the categories associated with the identified taxonomy, along with the number of entries associated with each of the categories associated with the identified taxonomies, will, in response to a string query, provide those entries which both contain the string and are associated with the identified taxonomy. The string is preferably one member of the group consisting of text, image, and graphic.

The present invention can be either a network of computers or a single computer.

The present invention preferably comprises a cache which stores the returned results of the search engine for rapid retrieval.

Various preferred embodiments of the invention have been described in fulfillment of the various objects of the invention. It should be recognized that these embodiments are merely illustrative of the principles of the invention. Numerous modifications and adaptations thereof will be readily apparent to those skilled in the art without departing from the spirit and scope of the present invention.